Storage Briefing: Trends and IU

Mike Floyd
Kurt Seiffert
Craig Stewart
George Turner
Dennis Cromwell
Dave Hancock,
Kristy Kallback-Rose
Matt Link
Steve Simms
Troy Williams

Indiana University

PTI Technical Report PTI-TR13-003

August 2013

Citation:

Floyd, M., Seiffert, K., Stewart, C.A., Turner, G., Cromwell, D., Hancock, D., Kallback-Rose, K., Link, M., Simms, S., Williams, T. 2014. "Storage Briefing: Trends and IU." PTI Technical Report PTI-TR13-003, August 2013.
http://hdl.handle.net/2022/18626

INDIANA UNIVERSITY

UNIVERSITY INFORMATION
TECHNOLOGY SERVICES

Table of Contents

Tables

Figures

1. Introduction

Storage technology is in a period of accelerating change. Rapid developments in flash memory and cloud services are the most visible indicators. Uses of storage face growing regulatory and funding agency compliance, and a growing bundling between software systems and storage options.
This briefing presents information on industry, regulatory, and usage trends at Indiana University (IU), and comparisons of storage services with Purdue and other Committee on Institutional Cooperation (CIC) schools. The sections of this document are largely independent of each other, so readers may focus on areas of interest.

2. Storage technology trends

2.1. Solid State (Flash) Storage

Flash is a significant factor in storage. Flash storage is the dominant storage in the mobile sphere and is rapidly replacing hard drives with solid-state devices (SSD) on laptops. SSDs are showing up in enterprise data centers, though adoption has been slow due to the price. Flash technology tends to be three or more times as expensive as the equivalent hard disk capacity, but can deliver orders of magnitude more performance than hard disks at a fraction of the energy cost. Flash will likely replace spinning disk for the highest enterprise tier over the next 2-3 years.

What has hampered widespread adoption in the enterprise is the limited number of writes that can be performed on a flash memory chip. Vendors have been addressing this with increasing sophistication in managing dead blocks and building spare blocks into the storage device. This in turn has limited the capacity of individual SSDs. While it is common to find 3TB or even 4TB hard drives in the enterprise, SSDs are typically no larger than 512GB, and 1TB drives are just now becoming available.

SSDs are the dominant form of flash storage in the enterprise. SSDs incorporate abstractions and standards intended to deal with the latency of physical spinning disk and head movement. Many of these abstractions make no sense when accessing flash memory and result in useless overhead. Many flash vendors are bringing products to market that bypass the hard disk interfaces and are realizing even faster and more resilient solutions. The fastest spinning disks deliver around 200 Input/Output Operations per Second (IOPS). A high-end SSD using a SATA (Serial Advance Technology Attachment) interface will deliver up 90,000 IOPS. Flash storage that bypasses the SATA interface can achieve over a million IOPS. Table 1 below compares several storage media.

	Access Time	Bandwidth	Cost as Multiplier of Tape	IOPS
SSD	100 µs	600MB/s	36x	Up to 90,000
Fastest disk	2.9ms	140MB/s	12x	~200
Slowest disk	12ms	40MB/s	3x	~75-100
Tape	seconds	250MB/s	1x	N/A

Table 1. Speeds and feeds for current storage technologies.

2.2. Fibre Channel Versus Ethernet

For the block-based SAN (Storage Area Network) environment the Fibre Channel (FC) protocol and fiber

media have dominated the storage. This infrastructure includes FC-enabled storage controllers, a unique dedicated FC switch, and FC host bus adapters in each server connected. FC speeds are 2Gb, 4Gb, 8Gb, and 16Gb per second.

File-based NAS (Network Attached Storage) systems are designed to leverage existing traditional copper-based Ethernet infrastructure. Many enterprises establish dedicated Ethernet networks for this data workload; others mix data and traditional network traffic. Ethernet speeds are generally a combination of 1Gb and 10Gb per second in terms of server connections.

The long-term trend is away from Fibre Channel and toward Ethernet for all data workloads. Ethernet speeds of 40Gb and 100Gb are beginning to find their way into major data centers for core networking capabilities. Once the price of the corresponding components (switch ports and server cards) nears the current cost of 10Gb equipment the economics and capabilities will favor Ethernet, even for the most demanding data situations. Some FC data workloads will migrate to the iSCSI (Internet Small Computer System Interface)—a protocol long supported by the Ethernet infrastructure. Others will migrate to FCoE (Fibre Channel over Ethernet) or FCIP (Fibre Channel over IP), which are competing standards (Cisco versus most other vendors).

Although FC and Ethernet dominate traditional storage systems, low-latency InfiniBand is becoming the switched fabric for many higher-end storage solutions, especially in high performance computing environments.

At IU:

- Research Technologies uses Ethernet fabric/switching exclusively to support the file-based NAS environment for services such as the Scholarly Data Archive and the Research File System.
- The Data Capacitor uses an Infiniband fabric to connect to Big Red II, but it also employs Infiniband-Ethernet routers for all other HPC workloads.
- Enterprise Infrastructure uses Fibre Channel fabric/switching almost exclusively to support the block-based SAN environment for all structured or database-intensive applications such as Oncourse and Kuali.
- A recent exception to FC deployment is the use of the iSCSI protocol over Ethernet for data backups to our new Virtual Tape Library environment.

2.3. Cloud Storage

Cloud storage and storage-related services are a substantial component of the IT landscape. They provide an off-premises path to economies of scale in cost and reduced responsibility, in exchange for reduced control. Cloud solutions exist for myriad storage problems including long-term storage, data backup solutions, collaboration, and data management and movement. Cloud storage solutions are likely to become the predominant consumer option. Gartner anticipates that 1/3 of all consumer data will be stored in the cloud by 2016.1 Institutions will increasingly leverage cloud storage as part of their IT portfolio. The federal government has adopted the Federal Cloud Computing Strategy, a cloud-first approach to future IT initiatives.2

However, large institutions will continue to struggle with balancing the economic efficiencies of cloud storage against business realities such as contracting, regulatory compliance, liability, technical integration standards, and internal standards of security. The Federal Cloud Computing Strategy offers a comprehensive framework in which to evaluate cloud solutions.

---

[1] http://www.gartner.com/newsroom/id/2060215

[2] http://www.dhs.gov/sites/default/files/publications/digital-strategy/federal-cloud-computing-strategy.pdf

In contemplating cloud storage strategies, carefully consider the amount of data to be stored and its growth, especially in terms of an exit strategy. Cloud solutions are heavily dependent on network access and bandwidth. When the time comes to end the relationship with the cloud storage vendor, establishing a suitable environment for the data –whether in-house or alternative cloud vendor– will be challenging, complicated by the amount of time it will take just to move the data. This will be costly, even when the vendor termination is orderly. Crisis situations will exacerbate these problems. As Dan Reed said, "Having a scientist use your supercomputers is nice, but once you have their data you have their soul," and any institution that focuses on intellectual and artistic innovation should consider carefully who has practical possession of their critical data assets. There are also legal questions, such as redress when a business goes bankrupt and physical facilities lie outside the US, and when data implies granting a license to the facility owner (e.g. Google docs).

Many cloud vendors offer specific storage solutions and some try to cover a wide variety of solutions. No one solution solves all storage problems. Examples of cloud storage solutions by category include:

- **Enterprise-contracted or consumer-grade, end-user file storage** (example: Storage as a Service solutions such as Box, G-Drive, SkyDrive, or DropBox). Most have consumer and enterprise offerings.
- **Archive Platform** (example: Amazon Glacier)
- **Backup Platform** (example: EMC Mozy or CrashPlan)
- **Object Storage** (example: SDSC Cloud Storage; based on OpenStack Swift)
- **Hybrid Storage** (example: StorSimple; requires local device, extends to cloud)

A rapidly growing area of innovation is a hybrid approach with onsite devices that help manage the enterprise connection to the cloud. These can provide local caching of data, support placement or mirroring of data among several cloud providers, and even interconnect cloud provider resources with local storage resources. Some vendors such as Oxygen-Cloud provide cloud services but use the enterprise local storage for the data store.

2.4. Data Movement and File Systems

Data movement – distinct from data storage – is now becoming such a challenge that data movement services are emerging.

Globus Online is an innovative approach that moves data in a "cloud" (Software as a Service) model. Globus Online is a web-based cloud service that allows a user to easily copy or move data between any two endpoints. Usually endpoints are well-defined storage services such as IU's Data Capacitor and Scholarly Data Archive (SDA). However, users can install the client component for Globus Online, which enables a user's laptop or desktop to become an endpoint. Users can use any browser to view their data sets on any endpoint to which they are authorized. They can request the data be moved or copied between the endpoints with over-the-wire encryption and verification on the transfer. The transfer is done asynchronously, and the user is notified when the transfer completes or a non-recoverable error occurs. Globus can handle petabytes of data in a transfer that can last weeks. Globus will automatically deal with transfer restarts should there be temporary network or endpoint outages.

IU has signed a contract with Globus Online granting access to all IU users, and CI-Login is used for authentication. Globus Online has been integrated into the IU Cyberinfrastructure Gateway.3 IU will also establish the new Research File System as an endpoint, allowing researchers to easily move large amounts of data between all the major storage services. The Globus restful API will enable applications to move

[3] https://cybergateway.uits.iu.edu/iugateway/

data in workflows.

IU is a leader in the use of Lustre file mounts over long-distance networks for file movement, and is now using IBM's GPFS (General Parallel File System) software. Mounting these systems across disparate systems allows standard tools and applications to move data over campus-scale distances.

Lustre remains a leader in high-performance object store systems. The Lustre file system is used by over 60% of the world's top 100 most powerful supercomputers. It is an open-source file system that is fast, scalable, and open, but implementing and operating it requires significant effort and expertise. Lustre is the main workhorse for HPC data at IU. It continues to be the best performing open-source cluster file system. This performance trend should continue with flash-based metadata servers. IU receives contracts for development of Lustre and our expertise with Lustre significantly aids our grant competitiveness. IU has a contract for $250,000 from OpenSFS, a non-profit organization dedicated to supporting and developing Lustre, and for designing and implementing new security features. This is based on IU's earlier work with Lustre security (UID mapping and correcting potential hacking vulnerabilities).

GPFS has been implemented widely and provides excellent integration with the HPSS (High Performance Storage System) tape archival system. IBM GPFS licensing terms have been onerous in the past. IBM has changed its position and IU recently purchased a new license allowing IU to broadly use this file system as the core of IUFS (Indiana University File System) and the new RFS (Research File System).

HPSS continues to be the best tape archive system. New sites adopt it every year, many switching from SAM-QFS. Just this year, HPSS released new RAIT technology, allowing sites like NCSA to stripe very large files across as many as 10 tape drives at a time with parity assuring a single tape failure will not precipitate data loss.

2.5. Big Data

Big Data is one of today's most prominent IT buzzwords. Gartner anticipates Big Data driving as much as $34 billion in IT spending in 2013. With that much money at stake, vendors will often blur the critical aspects of Big Data. Big Data solutions are not just massive amounts of data. Big Data can be best defined by volume, velocity, and variety.4. Volume is one challenge, but the rate data is acquired or needs to be processed (velocity) complicates matters. Variety can pose another challenge when the data do not sit well within existing database schemas. (Some take into account additional Vs, such as veracity). This requires huge amounts of bandwidth and computational resources. Clustered file systems and massive data center facilities are the major approach to most industrial-scale-sector Big Data challenges.

Like parallel computing before it, many disciplines and problem domains are discovering significant advantages in using Big Data solutions, most notably in analytics. Apache Hadoop has become the dominant framework in this area. Companies such as IBM have invested heavily around this framework. Big Data best fits in the realm of data scientists who can be defined as clever and curious storytellers with deep technical expertise in their scientific discipline.5

MapReduce algorithms comprise an evolving area of research. The iterative MapReduce research, Twister (Judy Qui, Indiana University) is an example. Beth Plale's work leading the HathiTrust Research Center is IU's clearest leasing example of a major Big Data project.

2.6. Structured and Unstructured Data

In general the management of data falls into two broad categories: Structured and unstructured data.

[4] http://strata.oreilly.com/2012/01/what-is-big-data.html
[5] http://radar.oreilly.com/2011/09/building-data-science-teams.html

Structured data is largely database-driven using block-based storage area network (SAN) technology used for relational databases like Oracle and SQL Server. Unstructured data are the data on file servers, web servers, media servers, and other data organized by files or objects. Lower-end storage systems tend to merge these two technologies into a common solution. Higher-end storage systems tend to be purpose-built for particular solution, e.g.: a scale-out network-attached storage (NAS)..Gartner estimates that unstructured data accounts for 80% of the storage in a typical enterprise.

Large-scale unstructured storage is moving towards clustered solutions. Traditional NAS vendors have been moving to offer clustered NAS heads. Large research institutions have been adopting clustered file system solutions like IBM's General Parallel File System (GPFS) or Lustre. Object storage systems in general offer great benefits, especially for large data objects, but issues related to POSIX (**P**ortable **O**perating **S**ystem **I**nterface) compliance tend to limit their applicability.

Document management is a significant issue within unstructured data. Huge swaths of organizational knowledge are maintained in various reports, spreadsheets, presentations, diagrams, and other document formats. Document management systems provide support for versioning, attribution, indexing, security, and other related capabilities. As much of this information has moved into the web, some web content management systems have stepped into this role. IU has two document management systems, OnBase and KnowledgeLake. OnBase has a longer history and has been integrated in highly structured workflows. KnowledgeLake is considered more accessible and a better fit with existing departmental tools and resources. UITS is undertaking an examination to determine if both warrant continued support.

Deduplication has been a popular concept in unstructured storage of late. This is an effective solution for email systems where the same mail message or document may be stored in several if not thousands of email accounts. It tends to be less useful with effective data management systems, which have principles of consistency and accountability that tend to reduce duplication.

A great deal of unstructured data tends not to be accessed after a few weeks. Many NAS solutions take advantage of this with tiered disk solutions.

3. Regulatory and funding agency requirements

3.1. HIPAA (Health Insurance Portability and Accountability Act)

IU was the first major supercomputer center to have all of its cyberinfrastructure aligned with HIPAA. There is now at least one other – SDSC (Dan Diego Supercomputer Center). IU's leadership in ability to analyze ePHI (electronic Personal Health Information) has been a critical asset in competing for NIH funds. One of the challenges with HIPAA remains that it is essentially self-asserted; there is no federal guidance on what it means to be HIPAA compliant. An organization is HIPAA aligned, which to a first order approximation means that security risks and controls are well documented and defined to the satisfaction of an entity's legal counsel.

3.2. FISMA (Federal Information Security Management Act)

FISMA, unlike HIPAA, creates specific guidelines for information security. FISMA, enacted in 2002, requires each federal agency to create security programs for itself, subcontractors, and grant awardees. FISMA compliance is becoming increasingly common as a requirement for grant awards from NIH (National Institutes of Health) and for contracts from entities such as the State of Indiana.

3.3. Federal Initiatives and Responses Regarding Open Data

On 23 February 2013, the White House Office of Science and Technology Policy (OSTP) issued a memorandum titled, "Increasing Access to the Results of Federally Funded Scientific Research" (http://www.whitehouse.gov/sites/default/files/microsites/ostp/ostp_public_access_memo_2013.pdf). This memorandum states:

To achieve the Administration's commitment to increase access to federally funded published research and digital scientific data, Federal agencies investing in research and development must have clear and coordinated policies for increasing such access. . .

> *each agency plan for both scientific publications and digital scientific data must contain the following elements:*
>
> *a) a strategy for leveraging existing archives, where appropriate, and fostering public-private partnerships with scientific journals relevant to the agency's research;*
>
> *b) a strategy for improving the public's ability to locate and access digital data resulting from federally funded scientific research;*
>
> *c) an approach for optimizing search, archival, and dissemination features that encourages innovation in accessibility and interoperability, while ensuring long-term stewardship of the results of federally funded research ….*

The SHared Access to Research Ecosystem (SHARE) proposal was put forth with the endorsement of AAU, APLU, and ARL in response to the OSTP request. The proposal envisions a phased set of federated repositories at research universities. IU has serious concerns regarding the proposal, and IT Vice President Brad Wheeler was recently asked by ARL's president to serve on a taskforce to assess and improve it.

Complying with open data directives from granting agencies will be a significant challenge for IU. This is important in terms of complying with funding agency directives and maintaining public credibility of university research. The latter is made more critical by recent articles about replicability of scientific research published in major scholarly journals.

4. Trends and Usage at IU and Elsewhere

4.1. Sensitive and Critical Data

The protection of sensitive and critical data is an essential aspect of data management at Indiana University. Data must be managed according to policy and legal requirements. As IU expands its data offerings with services like IUFS, it must select and apply appropriate safeguards to various storage options.

An important decision is whether encryption-at-rest data protection is required or whether compensating controls such as strong physical security can reduce the need for encryption-at-rest and still meet IU policy needs as stated in IT-12 and in the policies and practices of the Committee of Data Stewards.

The university does not necessarily need to avoid cloud services for sensitive and critical data. Many services provide high levels of security and can meet university requirements. In some cases, such as payment card processing, IU is required to use service providers who specialize in managing that data. Necessary contractual agreements (e.g., IU Data Security Addendum, HIPAA Business Associate Agreement) must be in place to enable services for sensitive and critical data.

4.2. IU Response to NSF Requirements for Data Management Plans

IU Libraries delivers and supports the Data Management Plan Tool software (DMPTool, https://dmp.cdlib.org/) that can automatically generate data plans, and contains customized information about IU's institutional data management resources. UITS and the IU Pervasive Technology Institute provide guidance on creating data management plans based on recommendations from a task force led by IU School of Informatics and Computing Professor Beth Plale.[6]

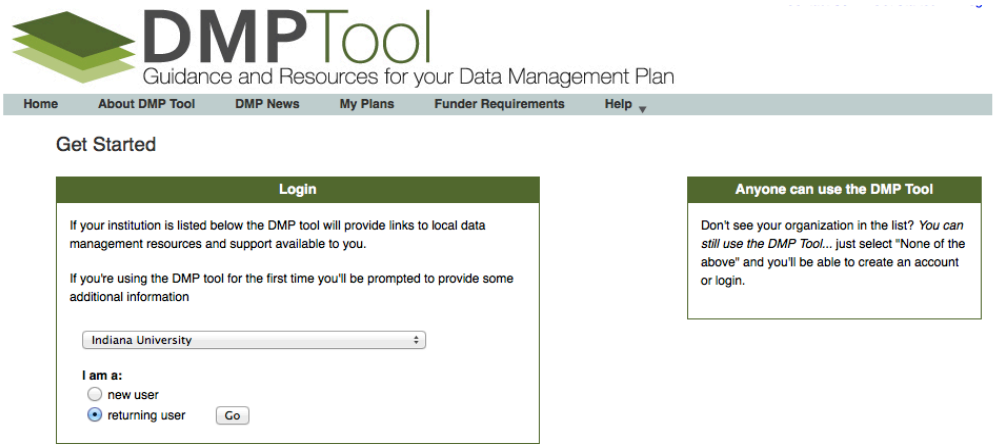

Figure 1. DMPTool login screen. Members of the IU community can log in and create customized data management plans.

4.3. Data Replication, Operation Continuity, and Tape-based Storage Systems

IU has made excellent use of its investment in data centers in Indianapolis and Bloomington. Both facilities, connected by a dedicated high-speed, low-latency network, form an excellent framework for data replication. This replication ensures protection against disasters and preserves operational continuity. Many applications and services could not afford such capabilities if data replication were not part of the core cyberinfrastructure. IU capability in this area is fairly unique among research universities.

The Tivoli Storage Manager (TSM) backup service replicates all backups between the data centers. This ensures recovery capabilities while mitigating the risk of data exposure by sending media off site. The UITS Enterprise Infrastructure (EI) division runs TSM for all database backups. As well, TSM is used for many system-level backups for servers located in the Informatics & Communication Technology Complex (ICTC) or Bloomington Data Center (BLDC). UITS operates a small TS3500 tape library at each data center to support these backups, along with newer (disk-based) Virtual Tape Libraries, which are the stated direction for supporting all future backups. This is a cost-recovery service.

UITS does not provide or support a backup solution for workstations or lab- and departmentally-based systems. Some users create system tarballs and copy them to the Scholarly Data Archive (SDA), or use other approaches to backing up data to SDA. Fundamentally, though, SDA is best suited for unstructured data.

Since 2006, data stored on the SDA have been by default replicated between IUB and IUPUI with the exception being data from the Large Hadron Collider. The IU copy of the collider data is at least the 3rd copy of any data – with an original copy at CERN and a duplicate copy at a US Tier 2 site). Whether an

[6] http://www.libraries.iub.edu/secure/defiles/NSF_DMP_Boilerplate_IUB-IUPUI_Fall_2012.doc

individual tape fails or an entire automated tape library is compromised, data archived in the SDA remains secure. The Enterprise SAN environment replication ensures that data is always available. It also ensures the operational continuity of the dependent applications and systems. RFS will also replicate data so that users can reliably access their data.

While the 50 miles between Indianapolis and Bloomington offers a great deal of protection against large catastrophes, events like Hurricane Sandy and Katrina remind us that even 50 miles can be too close to maintain operations in case of a regional disaster. In 2012 Indiana University and Texas Advanced Computing Center (TACC) signed a memorandum of understanding (MOU) allowing TACC to store up to 500TB in IU's SDA, and IU to store the same amount of data in TACC's equivalent archive system. This allows IU to replicate data sets that warrant out-of-region protection.

Figure 2 below shows UITS overall growth of tape capacity. While Gartner anticipates storage costs will decline by 35-45% per year, it also expects unstructured data will grow by 50% per year (+650% over the last five years). IU has seen 52% annual growth over the last five years in the Scholarly Data Archive, which is IU's single largest storage for unstructured data. Past and projected data growth stored in the SDA is shown below in Figure 3. SDA uses nearly 10.6PB of storage for 71 million files.

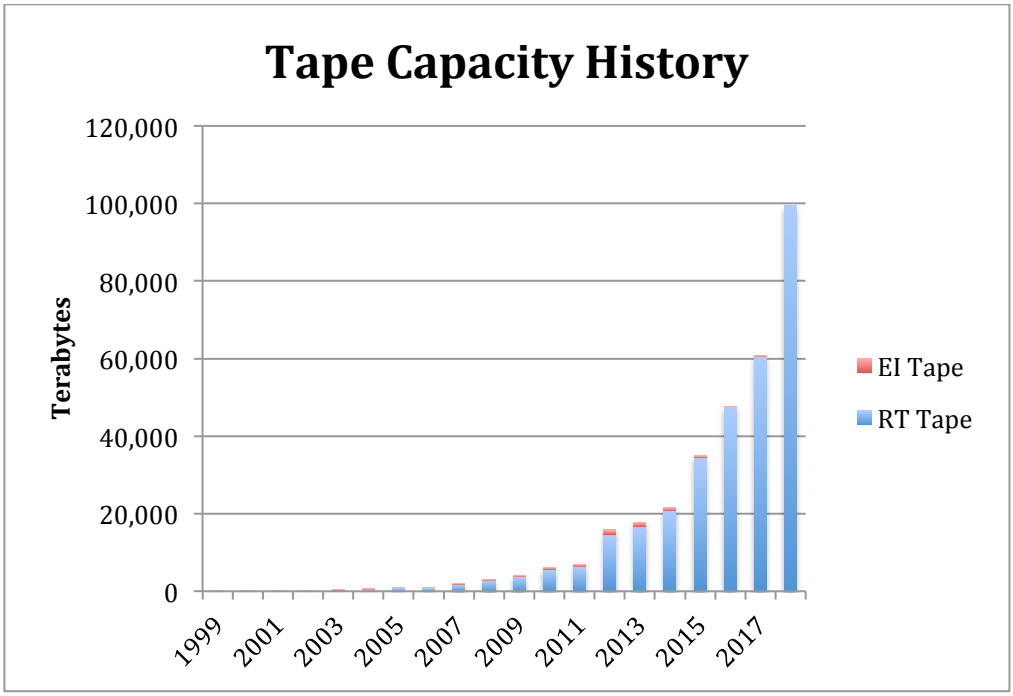

Figure 2. Overall growth of tape storage capacity for Research Technologies (RT) and Enterprise Infrastructure (EI) tape systems – actual until 2013 and projected after that.

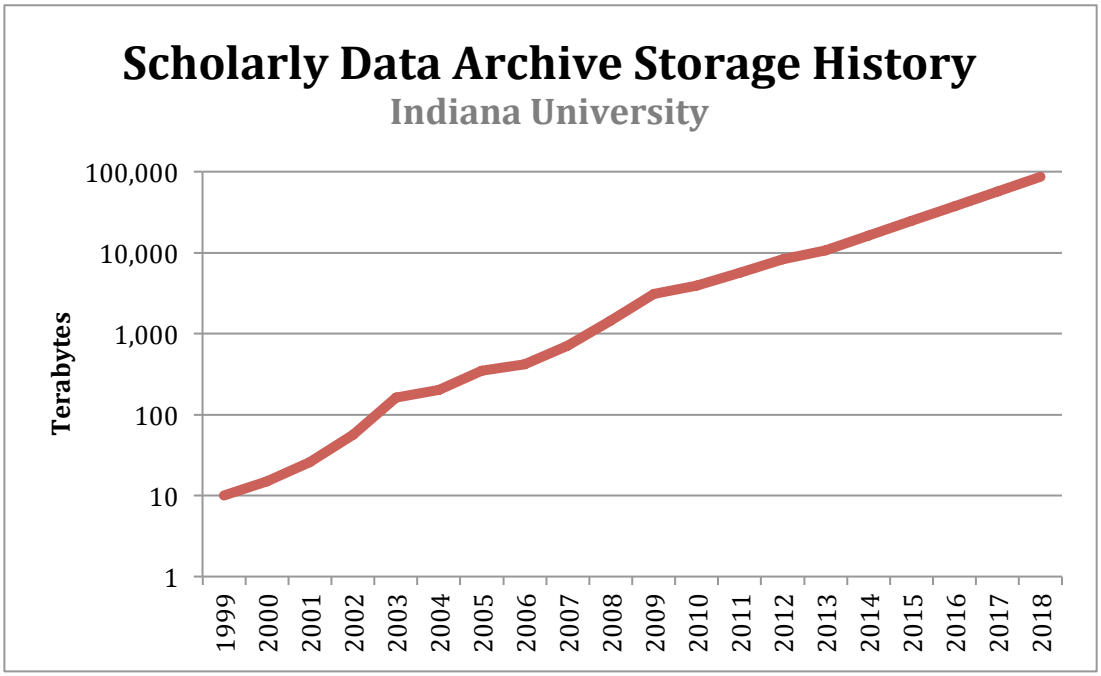

Figure 3. Growth of Data Storage in IU Scholarly Data Archive – IU's HPSS-based tape archive – actual until 2013 and projected after that.

4.4. Structured Data on Disk-based Storage Systems

Most of UITS structured storage resides on disk in the Hitachi SAN system. Enterprise database applications use this storage. RT maintains a much smaller pool of block storage for research databases in the Research Database Complex (RDC).

The Hitachi NAS system supports several storage services like the Consolidated Hosting Environment (CHE), file storage for Oncourse, SharePoint file storage, Slashtmp, and departmental file servers. While these services use file systems, the file systems are mostly built on logical unit numbers offered as block devices from the SAN. This results in relatively expensive file system storage.

Figure 4 shows growth in UITS disk storage systems, where the bulk of the storage is now unstructured data on the Data Capacitor/DCII.

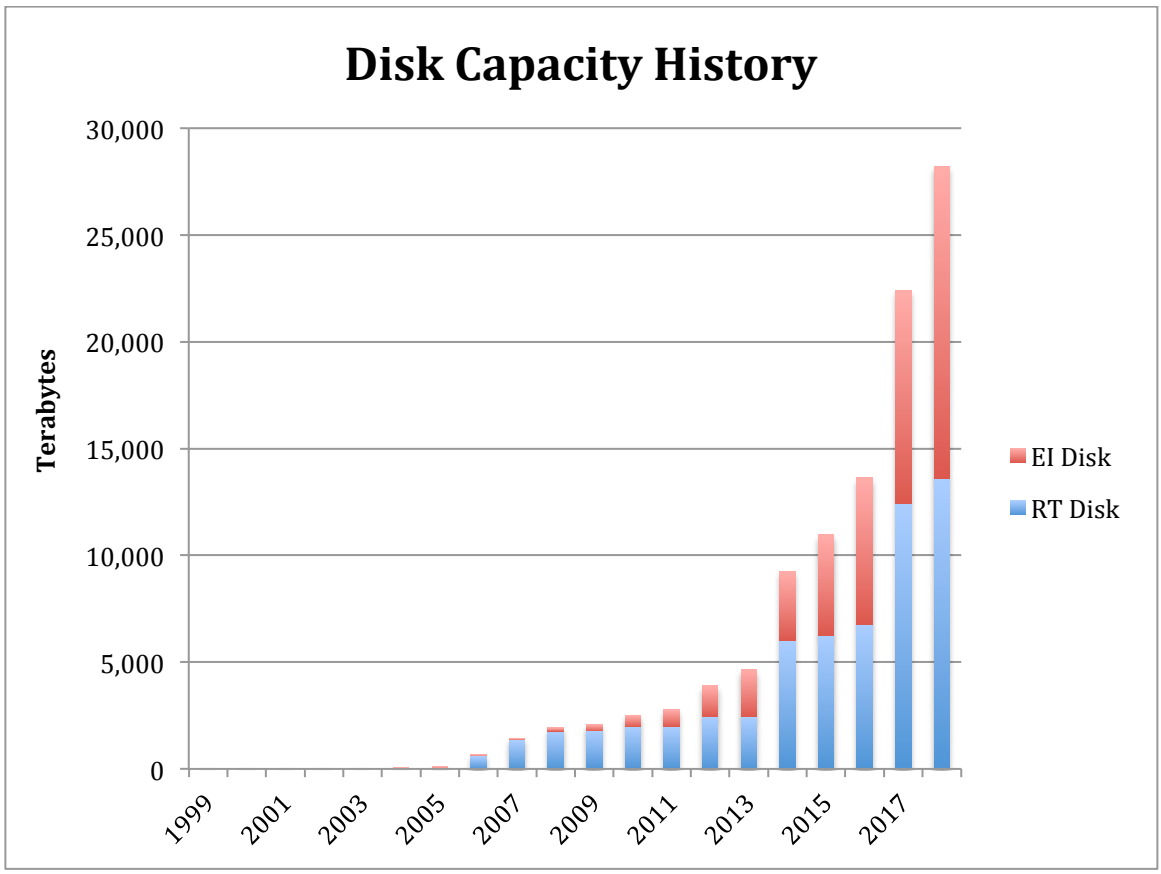

Figure 4. Growth in IU disk storage capacity for Enterprise Infrastructure and Research Technologies – actual until 2013 and projected after that.

4.5. Box@IU – contracted cloud storage service

In April 2012 IU signed an enterprise contract with Internet2's Net+ service for Box.com at a cost of $287,500 per year. IU played a major role in shaping the contract terms and service offering to use InCommon for identity. The Box terms allow up to 150,000 IU users to store up to 300TB collectively. To date 37,049 users have created Box@IU accounts, each with a 50GB quota. Box allows the user to store data in the cloud, but will also sync files between devices. It has extensive tools for versioning, collaboration comments, and managing group/public access. Box support for mobile devices extends collaborations to the environments most convenient for the users. Box is approved for administrative work at IU, unlike consumer-grade services where IU has no contractual assurances (e.g., DropBox).

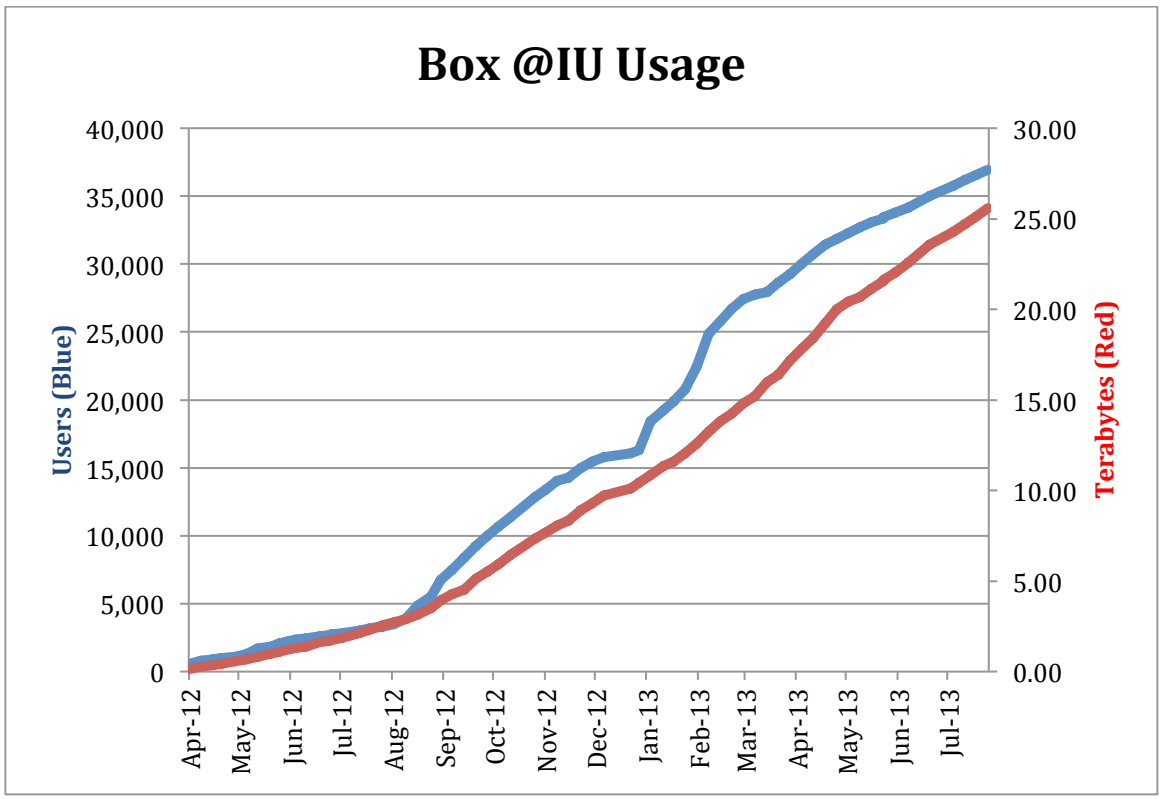

Figure 5. Growth in use of Box@IU.

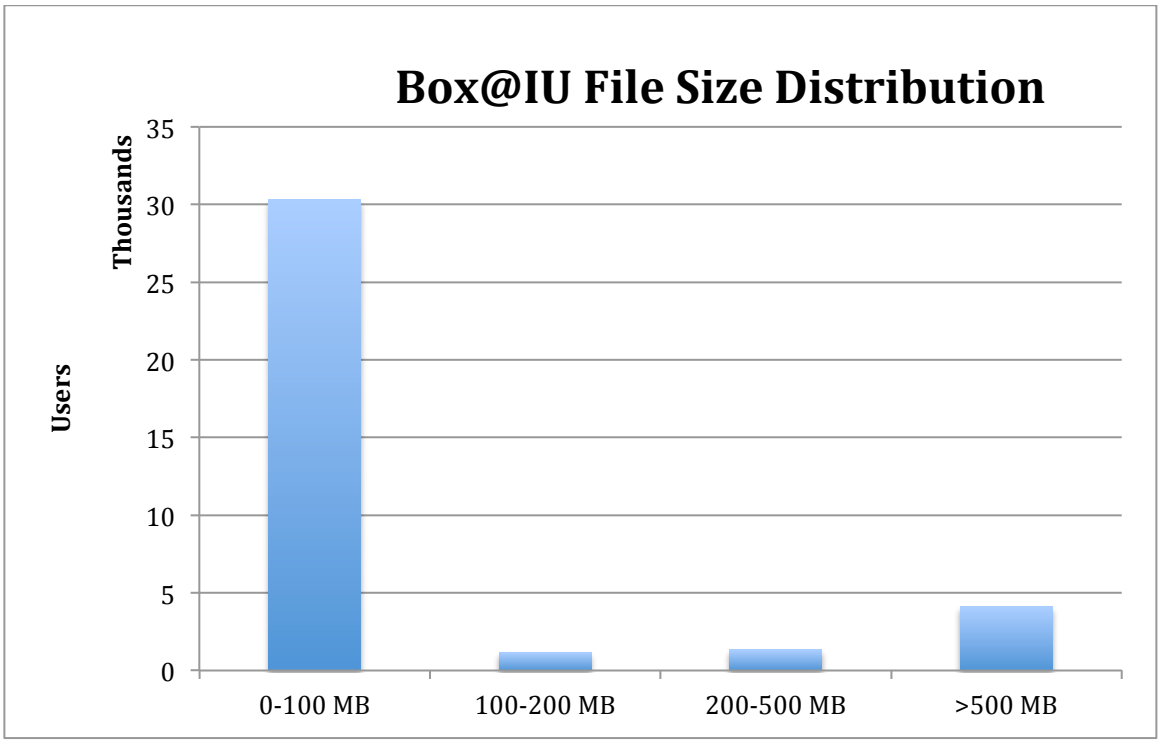

Figure 6. Distribution of file sizes stored on Box@IU.

The key value of Box as a storage system is its facile support of on-the-fly collaboration, within and beyond the IU namespace. The ability to share Box folders with other non-IU Box users is exceptionally

valuable for inter-institution collaboration. To date, IU users have collaborated with 4,630 non-IU users of Box. Most IU Box users do not store large amounts of data, as shown in the chart below. Box folders are often created, shared, then deleted for group tasks that may last only a few days or weeks.

Many universities have adopted Box. Box recently announced that over 100 universities have signed up for the Net+ service. The chart below compares Box usage from several US universities; all were early adopters. Box is not currently suitable for restricted or critical data.

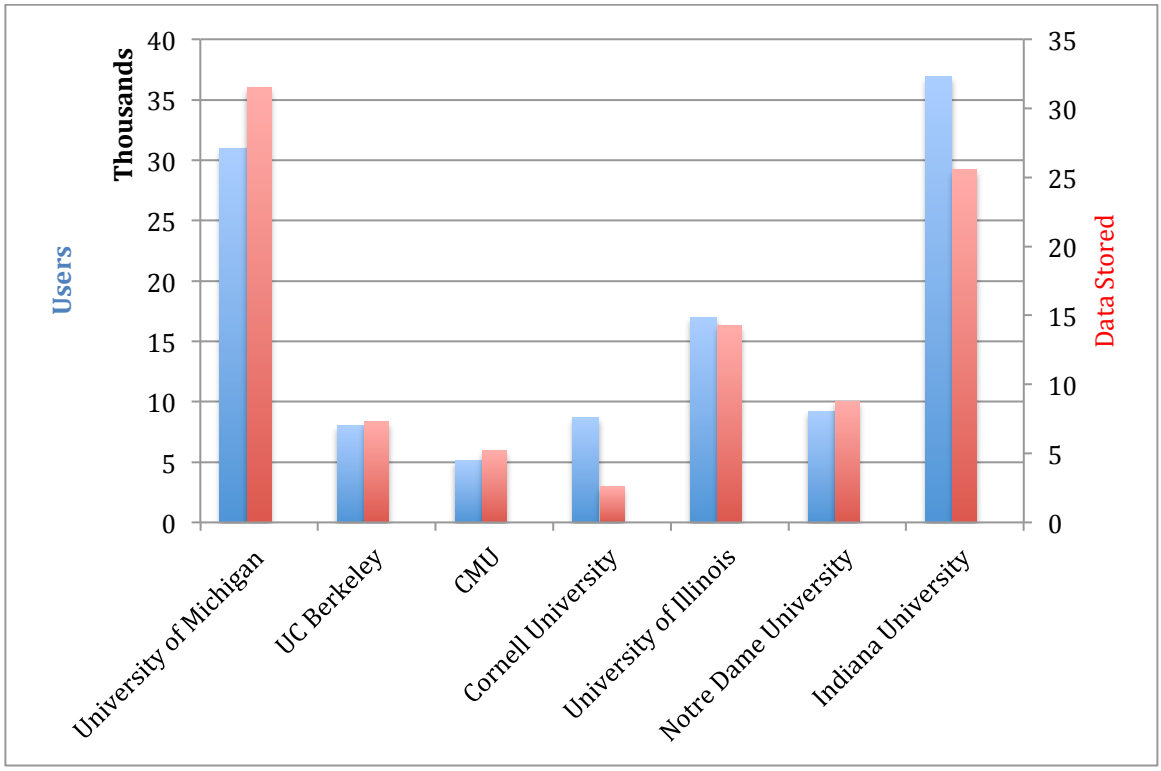

Figure 7. Box usage at a variety of leading universities.

4.6. Data Storage Applications

IU uses multiple data storage applications for a variety of types of data storage tasks. Table 2 lists the most heavily used applications by function:

Storage Application	Most Common Function
Temporary storage / Collaboration	
Box	Collaboration, file sharing, editing
SharePoint	Semi-permanent storage, collaboration
OnBase	Document management application used by many IU departments; will be infrastructure component in SIS.
KnowledgeLake	Document management application; built upon SharePoint

Alfresco Share	HIPAA-aligned collaboration tool
REDCap	HIPAA-aligned survey data collection, analysis, collaboration
Slashtmp	Short-term file exchange (HIPAA-aligned option)
Digital Asset Mgmt	PAGR digital asset repository; Migrating to SaaS Cloud
Teaching and learning	
Oncourse	
RDMS	
MySQL	
Oracle	
Clinical records systems for clinical/translational research	
Remedy	Registry system
Remedy-Informatic	Bio bank system
Forte OnCore	Clinical trials management system
Faculty publication / citation records	
VIVO	Grant-funded; at somewhat of a crossroads in terms of future. IU development led by Katy Boerner
Faculty Annual Review (FAR) System	Home-grown system that tracks 4,500 faculty across all IU campuses. Anne Massey, Kelley School of Business
PIVOT	Thompson Reuters product
ReSEARCH	Elsevier / Sci Vol product
Preparation of Data Management Plans	
DMPTool	Tool supported by IU Libraries for template-driven preparation of data management plans used in preparing grant proposals and executing grant awards
Code repositories	
Subversion	Longstanding code management/repository
Jira	Complex, sophisticated-users-only code management
Github	Code management / repository, collaboration, project management
Curated data archives	

IUScholarWorks (Dspace)	
Fedora	

Table 2. List of data management and access applications in wide use at IU.

5. Comparisons – IU Solutions and Others

5.1. Cost of Solutions Used by IU Community

Figure 8 compares the storage costs per GB for several file storage systems used at IU.

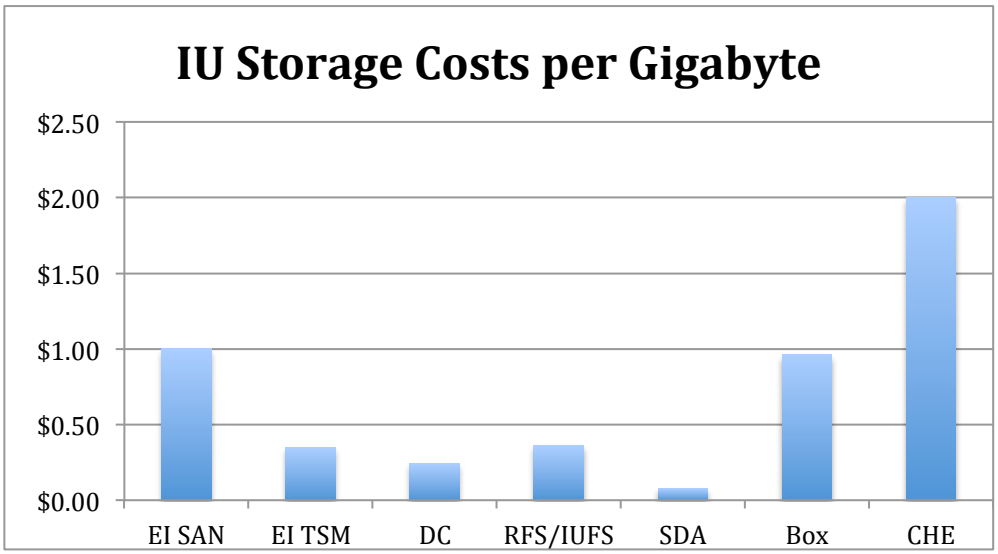

Figure 8. Comparison of costs of different file storage systems in use at IU. Abbreviations expand to: Enterprise Infrastructure SAN/TSM, Data Capacitor, Research File System, Scholarly Data Archive, Common Hosting Environment (Windows-based).

At $0.96/GB/year, Box is as expensive as IU's Enterprise Infrastructure (EI) SAN solution and nearly three times more expensive than the Research File System.

Storage system	Price/TB	Total TB stored	# Users	# Files
Box	958	26	37,049	
SharePoint	1,024	2.8	3,727	
RFS (GPFS)	369	89.6	2,856	80,169,271
Data Capacitor II	122	1,345	12,382	239,171,506
SAN - EI Hitachi (structured)	1,024	2202	N/A	N/A
SDA (HPSS)	81.9	10,620	2,603	71,380,997
TSM	256	1242	1,642	1,603,078,109

			servers	
Amazon Glacier (Net+ pricing)	123	N/A	N/A	N/A

Table 3. Comparison of different storage systems in terms of price, usage, and performance. For DCII and RFS, these data show projections of metrics for when migration to new DCII and RFS is complete.

5.2. Purdue

5.2.1. Purdue and EMC

In June 2012 Purdue and EMC Corporation announced a five-year relationship to address directly accessible storage for students and faculty. The Purdue-EMC announcement also listed the following collaboration objectives:

- EMC and Purdue will jointly develop a new architecture for research data management and curation.
- EMC and Purdue will develop new technologies to ingest, analyze, transfer, and manage enormous research data sets, especially in the field of bioinformatics.
- With EMC's assistance, Purdue will begin building a repository of all Purdue intellectual property.

Purdue does not currently offer a unified namespace for their storage services, and by default most academic user data is kept as single copy only.

5.2.2. PURR (Purdue University Research Repository)

PURR is a customized implementation that uses Purdue's HUBzero collaboration tool (http://hubzero.org/). IU makes extensive use of the HUBZero platform for the Center for Translational Sciences Institute. The PURR use is a customized product, not just another HUB. PURR's ISO certification is based on procedures, policies, and documentation, and is an important part of this product. It uses the tools and resources shown in Figure 2.

Figure 9. Structure of PURR (from a HUBZero teleconference).

PURR allows easy publishing of small data sets through the automatic creation of DOIs and Excel-like browsing capabilities. PURR has an easy-to-use interface with dataset publication tools supported by an ISO-compliant process. In that regard, it is an innovative development on the HUBzero platform.

5.3. CIC Schools

Based on a recent survey of Committee on Institutional Cooperation (CIC) schools done by the Storage working group, here are some reference points related to personal storage. All 13 CIC schools responded.

All schools except Nebraska-Lincoln provide some sort of base-funded university storage. The technologies supporting these services include traditional Common Internet File System (CIFS) file shares, Samba, Box.com, Oxygen-Cloud, Google Apps, and Andrew File System (AFS). Base storage allocations per person ranged from a low of 1GB (Northwestern, University of Chicago, and University of Wisconsin-Madison) to a high of 50GB (IU), with a typical allocation around 10GB.

For research-specific storage, all schools except Nebraska-Lincoln provide centrally managed services ranging in size from 50GB to hundreds of terabytes, in varying forms of base-funded and charge-back models.

All schools except University of Chicago, University of Illinois at Urbana-Champaign, Minnesota, Nebraska-Lincoln, and Northwestern provide some form of institutional data repository or repositories.

Only Iowa, Northwestern, and Purdue reported a formal integrated data curation service.